Scaling of MOS technology

Carver A. Mead, California Institute of Technology

The MOS transistor is the workhorse of modern microelectronics. Reducing the feature size of CMOS fabrication processes has been the primary method by which ever-increasing computation could proceed at ever-decreasing cost and power consumption. How does this scaling affect device performance? Are there fundamental physical limits to how small the MOS device, as we know it today, can be scaled?

Transistor current is the flow of mobile channel charge induced by an equal charge on the gate. For a logic circuit, the supply voltage induces the channel charge, creates an electric field in the channel, and is the difference between output logic levels.

In long-channel devices, the charge velocity is proportional to the electric field in the channel. The channel current is the product of the channel charge and velocity. Therefore, the device current has a quadratic dependence on the supply voltage. This current must charge the load capacitance to approximately one-half of the supply voltage to achieve a logic transition. Thus, circuit speed is linear in the supply voltage—a dependence that kept power-supply voltages artificially high until a few years ago.

For device dimensions below 1 micron, the old scaling dependence no longer holds—charge velocity becomes independent of electric field. Decreasing the supply voltage no longer decreases the channel current. The same factor decreases both the output current and output voltage. In this regime, the only effect of decreased supply voltage is a decrease in the switching energy, with virtually no decrease in performance.

It is imperative to reduce the supply voltage for reasons other than reducing power consumption. To induce sufficient charge in the channel with a lower operating voltage, we must further thin the gate oxide. The sum of the source and drain depletion-layer thicknesses must be less than the channel length. It is inevitable that, as "minification" continues, these dimensions will become sufficiently small that electron tunneling through them will become comparable to other device currents. These parasitic currents are exponential functions of the supply voltage.

I have presented these considerations earlier.[1] The most remarkable conclusion of my work is that transistors with 0.03-micron channel lengths will operate on a 0.4-volt power supply about three times faster than do today's best devices. Only below this scale do parasitic currents overwhelm the energy consumed in the performance of real computation.

The enormous effect of device scaling on computational capability becomes apparent only when viewed from the system level. We'll see systems integrated to upward of 10^9 devices per square centimeter. Interconnects—both within a single chip and across chip boundaries—determine the dominant signal latency. Even today, it has become more economical to break each chip into several processors that can operate in parallel than to build larger "dinosaur" processors.

Massive parallelism is possible in present-day technology; it will become mandatory if we are to realize even a fraction of the potential of more highly evolved technology. Each processor can operate with its own local synchronous timing, with self-timed signaling between processors.

We have never been able to see more than about two technology generations ahead. In spite of our myopia, *the technology will continue to evolve.* It will evolve because that evolution is possible; because we gain so much at the system level by that evolution; and because the same energy and will on the part of bright, energetic, devoted people that have overcome enormous obstacles in the past will overcome those that lie ahead.

***Carver A. Mead** is the Gordon and Betty Moore professor of engineering and applied science at the California Institute of Technology in Pasadena, California. He works on VLSI design, neuromorphic systems, and the physics of computation.*

Reference

1. C. Mead, "Scaling of MOS Technology to Submicrometer Feature Sizes," *J. VLSI Signal Processing*, Vol. 8, 1994, pp. 9-25.

We will see clock speeds of about 900 MHz with a 60 ISPEC95 rating in 2000. Such tremendous clock rates place great demands on the resistance and capacitance of the chip's metal interconnects for power and clock distribution. These multimillion-transistor devices also face new hurdles in packaging and power management.

Architecture. In the late 1980s, there was much debate about which microprocessor architecture held the key to fastest performance. RISC (reduced instruction set computing) advocates boasted faster speeds, cheaper manufacturing costs, and easiest implementation. CISC (complex instruction set computing) defenders argued that their technology provided software compatibility, compact code size, and future RISC-matching performance.

Today, the architecture debate has pretty much become a nonissue. Both the debate and the competition have been good for the industry, as both sides learned a great deal from the other, which stimulated faster innovation. There is really no perceptible difference between the two in either performance or cost. Pure RISC chips like the IBM ROMP, Intel 80860, and early Sun Sparc, as well as pure CISC chips like the DEC VAX, Intel 80286, and Motorola 6800, are gone. Smart chip architects and designers have incorporated the best ideas from both camps into today's designs, obliterating the differ-